(19) Europäisches Patentamt
European Patent Office
Office européen des brevets

(11) EP 0 884 876 A1

(12) EUROPEAN PATENT APPLICATION

(43) Date of publication:
16.12.1998 Bulletin 1998/51

(51) Int. Cl.[6]: **H04L 12/56**, H04Q 11/04

(21) Application number: **98201768.3**

(22) Date of filing: **25.05.1998**

(84) Designated Contracting States:
AT BE CH CY DE DK ES FI FR GB GR IE IT LI LU MC NL PT SE
Designated Extension States:
AL LT LV MK RO SI

(30) Priority: **31.05.1997 US 48193 P**

(71) Applicant:
TEXAS INSTRUMENTS INCORPORATED
Dallas, TX 75265 (US)

(72) Inventors:
- **Ren, Jing-Fei**
  Plano, TX 75025 (US)
- **Landry, Randall J.**
  St. Merrimack, NH 03054 (US)
- **Izzard, Martin John**
  Dallas, TX 75248 (US)

(74) Representative: **Holt, Michael**
Texas Instruments Ltd.,
PO Box 5069
Northampton, Northamptonshire NN4 7ZE (GB)

(54) Improvement packet switching

(57) A packet switch has N digital input ports (28) of bandwidth B for receiving data cells including destination addresses for determining output ports, a shared input cache (32), N memory modules of bandwidth N • B for buffering, a switch fabric, and N digital output ports. The digital multiplexer (30) receives each data cell from the input ports and writes it to the shared input cache together with a corresponding port queue number, queue position, & memory module number in response to its destination address so that (1) cells having the same queue number are cyclically assigned to different memory modules and (2) cells having the same queue position are cyclically assigned to different memory modules. The digital demultiplexer (34) reads each data cell from the shared input cache and writes it to one of the N memory modules according to its assigned memory module number and queue position. Then the switch fabric reads the data cells in each memory module by queue position and writes each to a corresponding output port matching the cell's queue number.

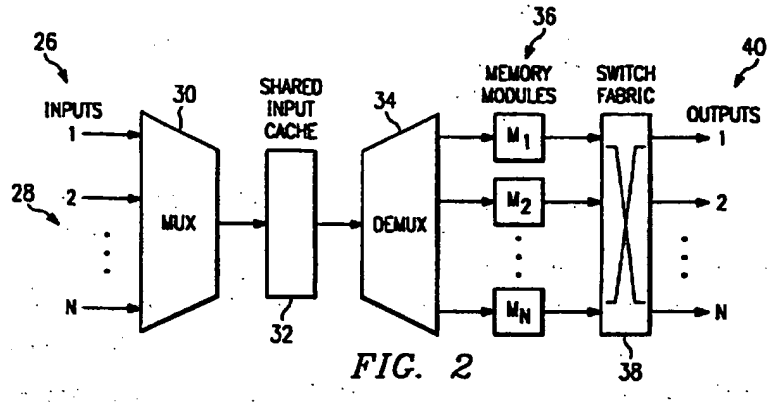

FIG. 2

EP 0 884 876 A1

Description

FIELD OF INVENTION

Our invention relates to packet switching, and more particularly to architectures for, and methods of using, multiport packet switches, particularly at high speeds such as gigabits/second with good delay-throughput.

BACKGROUND OF THE INVENTION

Conventional shared-memory packet switch architecture makes best use of memory capacity while achieving the optimal delay-throughput properties. However, for N ports, the shared-memory's bandwidth has to be N times each individual port's bandwidth B. For a multiport gigabit packet switch, this requires using expensive fast SRAM and wide memory interfaces for a multiport gigabit packet switch.

Researchers have been working on building fast switches out of memory modules operating at port speeds. For example, an input-queuing switch architecture uses N memory modules of bandwidth B, one for each port. But the basic input-queuing architecture suffers from head-of-line blocking and only achieves about 63% throughput. Although sophisticated scheduling algorithms have been proposed to improve the performance of the input-queuing switches, they have yet to achieve the ideal delay-throughput properties and efficient memory capacity utilization of shared-memory architecture.

Another approach is a shared-multiple-memory module (SMMM) architecture independently proposed by (1) H. Kondoh, H. Notani, and H. Yamanaka of Mitsubishi Electric Corp. in A Shared Multibuffer Architecture for High-Speed ATM Switch LSIs, IEICE Trans. Electron. Vol.E76-C, No.7, July 1993, pp.1094-1101, and S. Wei and V. Kumar of AT&T Bell Labs in (2) On the Multiple Shared Memory Module Approach to ATM Switching, Proceedings of IEEE ICC 1992, pp. 116-23 and (3) Decentralized Control of a Multiple Shared Memory Module ATM Switch, Proceedings of IEEE ICC 1992, pp.704-708, 1992, each of which articles is hereby incorporated by reference.

For SMMM the N input ports are connected to M memory modules which are in turn connected to the N output ports, conceptually through two switch fabrics. Although Bell Labs' switch architecture using either a centralized scheduling scheme in (2) or a decentralized scheduling scheme in (3) can provide an ideal delay-throughput, it requires 2N - 1 memory modules, each having a bandwidth B. The Mitsubishi Electric switch requires N memory modules of bandwidth 2B to achieve a reasonable throughput. And while memory architecture has been extensively studied in the context of the multiple processing, because packet switching has a different ordering, that earlier architecture cannot be used for a switch design.

Although for many years memory cell capacity has been increasing exponentially, memory bandwidth has only been increasing linearly. So one object of our invention is to build an N-port switch that only requires N memory modules of bandwidth B, as an input-queuing switch does. This would make it possible to build the fastest switch with a given memory technology or build switches with inexpensive RAMs. Other objects of our invention are to achieve optimal delay-throughput to meet performance requirements and to allow maximum sharing of memory space to enable the switch product to be competitively priced.

SUMMARY OF THE INVENTION

Our packet switch has a novel distributed shared-memory architecture providing N digital input ports of bandwidth B for receiving data cells including destination addresses for determining output ports, a shared input cache, N memory modules of bandwidth N • B for buffering, a switch fabric, and N digital output ports. A digital multiplexer 30 receives each data cell from the input ports and writes it to the shared input cache together with a corresponding port queue number, queue position, & memory module number in response to its destination address so that (i) cells having the same queue number are cyclically assigned to different memory modules and (2) cells having the same queue position are cyclically assigned to different memory modules. Next a digital demultiplexer 34 reads each data cell from the shared input cache and writes it to one of the N memory modules according to its assigned memory module number and queue position. Then the switch fabric reads the data cells in each memory module by queue position and writes each to a corresponding output port matching the cell's queue number.

Our invention also includes a new method of operating a packet switch having N digital input ports of bandwidth B for receiving data cells including destination addresses for determining output ports, a shared input cache, N memory modules of bandwidth N • B for buffering, a switch fabric, and N digital output ports. In our method, first each data cell received by the ports is written to the shared input cache together with a corresponding port queue number, queue position, & memory module number in response to its destination address so that (i) cells having the same queue number are cyclically assigned to different memory modules and (2) cells having the same queue position are cyclically assigned to different memory modules. Next each data cell is read from the shared input cache and written to one of the N memory modules according to its assigned memory module number and queue position. Then the data cells in each memory module are read by queue position and each written to a corresponding output port matching the cell's queue number.

Our distributed shared-memory architecture uses only a small input cache and N memory modules of

bandwidth B to implement an N-port packet switch, its aggregate memory bandwidth is only N • B. While the architecture has the same memory bandwidth requirement as the input-queuing, it achieves virtually the ideal delay-throughput performance and maximum memory capacity utilization as the shared-memory switch. We believe this architecture is particularly suitable to low cost multiple port gigabit switches using commercial DRAM modules. These and further advantages of our invention will become more apparent by way of example in the detailed description below.

BRIEF DESCRIPTION OF THE DRAWINGS

The present invention will now be further described, by way of example, with reference to the accompanying drawings in which:

Fig. 1 is a general sketch of a packet switch 10 for switching 14 data cells arriving at N input ports 12 into cells 16 directed to N output ports 20.

Fig. 2 is a block diagram of an embodiment of a distributed shared-memory switch according to our invention.

Fig. 3 illustrates how N logical queues, one for each output port, are two-dimensionally distributed in the memory modules of Fig. 2.

Fig. 4 is a block diagram of an event-driven simulation model at the node level of a 16 port distributed shared-memory switch using an Opnet modeler to study performance of the switch architecture.

Fig. 5A shows the cell input rate (cells/sec) measured at input port in_0,

Fig. 5B shows the output rate measured at output port out_0,

Fig. 5C shows the total number of cells in the input cache, and

Fig. 5D shows the number of cells in memory module mm_0 in the simulation of Fig. 4, for the simulation model of Fig. 4.

DETAILED DESCRIPTION

An embodiment of our distributed shared-memory switch 26 is shown in Fig. 2. Switch 26 has N input ports 28 coupled by a digital multiplexer (MUX) 30 to a shared input cache 32, a digital demultiplexer (DEMUX) 34, N memory modules 36 and a switch fabric 38 coupling memory modules 36 to N output ports 40. We will use the generic term "cell" to refer to a data segment of fixed length handled by the switch. The transmission time of a cell at the port speed is measured in slot time. Since the memory modules operate at the same speed (or bandwidth) as the ports, at most one cell may be written into and at most may be read from a memory module per slot time.

While there are numerous ways to assign arriving cells to the N memory modules, we use a two-dimensional cyclic order paradigm from the output logical queues' perspective, as illustrated by Fig. 3 for a 4-port switch. Each output port has a corresponding logical queue. Cells in the same logical queue are buffered to resolve output conflicts and then sent out to a corresponding output port, one per slot time. Cells in the same logic output queue are placed in different memory modules in a cyclic fashion. Furthermore, cells belonging to different logic queues but the same queue position are placed in different memory modules in cyclic order.

Because only one cell may be written into a memory module in one slot time, it is not always possible to place all arriving cells in the memory modules in the two-dimensional cyclic order. Therefore, if more than one arriving cell is assigned to the same module to meet the cyclic order requirement, all the cells but one are temporarily buffered shared input cache 32. To enable sharing of cache 32, multiplexer 30 and demultiplexer 34 are used. When up to N cells arrive at the beginning of a slot,multiplexer 30 assigns each of its own memory module numbers following the two-dimensional cyclic order and sends them to shared input cache 32. The input cache is organized as N queues, one for each memory module. At each slot time, demultiplexer 34 routes the first cell (if any) in each queue to a specified memory module. The newly arriving cells join the tails of the queues according to their module number. We can show that if the cache memory is completely shared by all N queues and cells are assigned module numbers based on their arrival order, by following the two-dimensional cyclic distribution there are at most

$$\sum_{i=1}^{N-1} i = \frac{N(N-1)}{2}$$

cells in the cache. For example, for a switch of 16 ports, a cache of only 120 cells is sufficient. If each cell is 72 bytes in length, this cache requires less than 70 Kbits.

The N memory modules provide the large buffer space required for the fast packet switch. Since a double cyclic order is followed when placing cells into memory modules, at any slot time the switch fabric can read out the cell at the head of each logic queue from the memory and route it to its appropriate port. The simplicity and regularity of the two-dimensional cyclic order facilitates scheduling cell transmissions over the fabric.

To study our switch's performance, we used an Opnet modeler to build a event-driven simulation model 46 of a 16 port distributed shared-memory switch. Fig. 4 shows switch model 46 at Opnet's node level. The input 128 and output 140 ports are respectively denoted by in_i and out_i (i=0,1,...,15). The functions of multiplexer 30 and demultiplexer 34 were modeled by using a mechanism in the cache module 132 that handles multiple data streams. Cache module 132 also modeled

other functions required by the shared input cache 32, including queuing cells according to the memory module numbers and memory sharing among all the queues. The memory modules 136 were modeled by the (N=16) mm_i modules in the model. Finally, the switch fabric was modeled by the module called crossbar (CRBAR) 138. To meet the bandwidth constraint, each link was only allowed to transmit one cell during a slot time. Similarly, each memory module could only receive and transmit one cell at most during a slot time.

To assess switch performance, a cell generator was connected to each input port. One cell was generated per slot time, providing a 100% traffic load. The cell's destination was uniformly chosen from the other N - 1 ports. Therefore, the traffic was symmetric. The destination ports of successive cells from the same cell generator were correlated to create burstiness from the output ports' perspective. The burstiness was adjustable by a correlation parameter. Each output port was connected to a traffic sink where its received cells were destroyed.

Figs. 5A-5D show some results for a 0.5 second simulation run. Fig. 5A shows the cell input rate (cells/sec) measured at input port in_0. All the other 15 ports should have the same input rate. Fig. 5B shows the output rate measured at output port out_0. The output rate converges to the input rate, indicating that a 100% throughput is achieved (the dispersion between input and output rates is mostly due to output port conflict inherent in all packet switches with burst traffic).

Fig. 5C shows the total number of cells in the input cache. There is a range for the number of cells for a given time because the measurement is taken after arrival and departure of cells at the input cache. The lowest value is the number of cells after departure and the highest value is the number of cells after arrival. As expected, under 100% traffic load, the number of cells in the input cache increases monotonically, but it is well below the given bound 120 cells even after 0.5 second.

Finally, Fig. 5D shows the number of cells in memory module mm_0. Again, we see it increases monotonically. Queues are built up here due to output conflict.

Although the detailed embodiment and simulation described in this disclosure are for memory modules with the port bandwidth, our distributed shared-memory switch architecture can be easily extended to the case where memory modules are faster than port speed. For instance, a switch of N ports may be built out of N/2 modules of speed 2B.

Claims

1. A packet switch comprising:

   N digital input ports each having a bandwidth B for receiving a corresponding arriving stream of input data cells, each data cell including a destination address from which an output port can be determined;

   a shared input cache for storing the data cells received at the input ports;

   a digital multiplexer for receiving each data cell from the input ports and for writing each data cell to the shared input cache together with a corresponding port queue number, queue position, and memory module number in response to its destination address such that;

   (i) cells having the same queue number are cyclically assigned to different memory modules;

   (ii) cells having the same queue position are cyclically assigned to different memory modules;

   N memory modules each having a bandwidth N • B for buffering a stream of data cells;

   a digital demultiplexer for reading each data cell from the shared input cache and for writing each data cell to one of the N memory modules according to its assigned memory module number and queue position;

   N digital output ports each having a bandwidth B; and

   a switch fabric for reading the data cells in each memory module by queue position and for writing each data cell to a corresponding output port matching the cells queue number.

2. A method of operating packet switch including N digital input ports each having a bandwidth B, which method comprising:

   receiving an arriving stream of input data cells at the input ports, each data cell including a destination address from which an output port can be determined;

   storing the data cells received at the input ports in a shared input cache;

   writing each data cell to the shared input cache together with a corresponding port queue number, queue position, and memory module number in response to its destination address such that;

   (iii)cells having the same queue number are cyclically assigned to different memory modules;

   (iv) cells having the same queue position are cyclically assigned to different memory modules;

   buffering the stream of data cells with N memory modules each having a bandwidth N • B;

   reading each data cell from the shared input cache and writing each data cell to one of the N

memory modules according to its assigned memory module number and queue position; and

reading the data cells in each memory module by queue position and writing each data cell to a corresponding one of N digital output ports each having a bandwidth B, the output port matching the cells queue number.

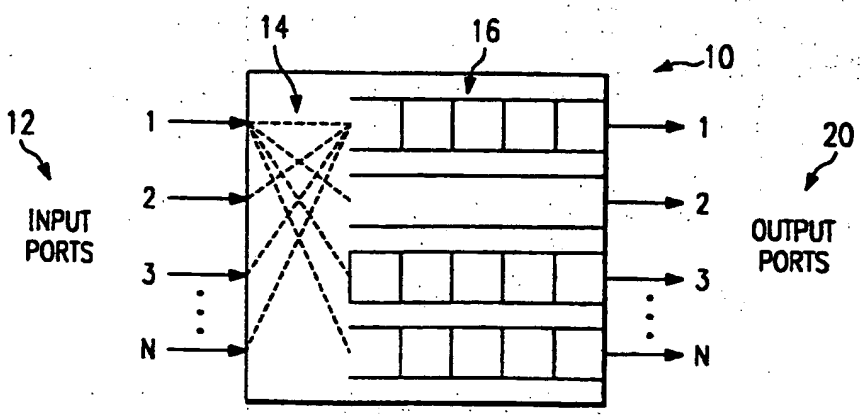

FIG. 1

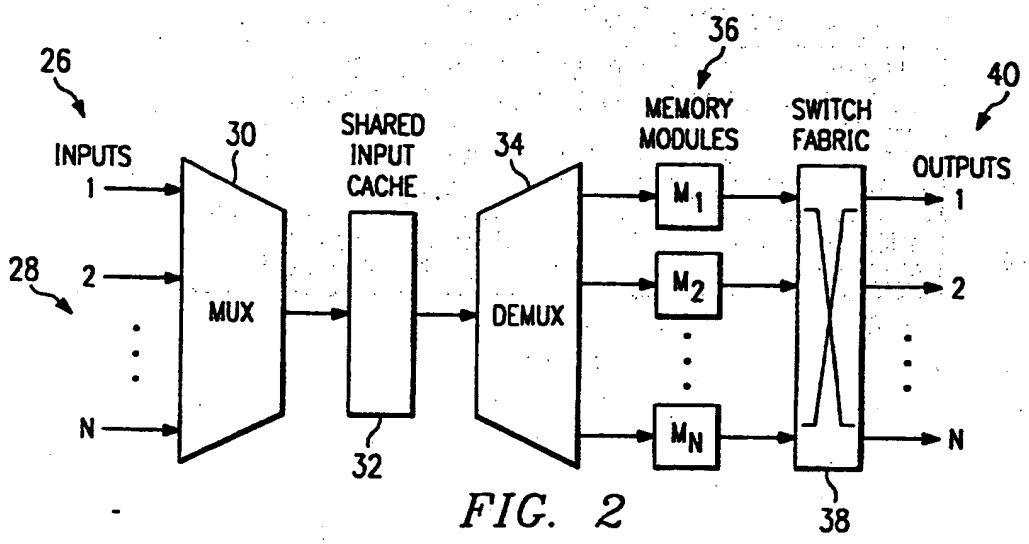

FIG. 2

LOGICAL QUEUES 2-DIMENSIONALLY DISTRIBUTED IN MEMORY MODULES (4-PORT SWITCH)

							OUTPUT PORT QUEUE
	MEMORY MODULE2	MEMORY MODULE1	MEMORY MODULE4	MEMORY MODULE3	MEMORY MODULE2	MEMORY MODULE1	1
		MEMORY MODULE2	MEMORY MODULE1	MEMORY MODULE4	MEMORY MODULE3	MEMORY MODULE2	2
			MEMORY MODULE2	MEMORY MODULE1	MEMORY MODULE4	MEMORY MODULE3	3
		MEMORY MODULE4	MEMORY MODULE3	MEMORY MODULE2	MEMORY MODULE1	MEMORY MODULE4	4
6	5	4	3	2	1	0	

QUEUE POSITION

FIG. 3

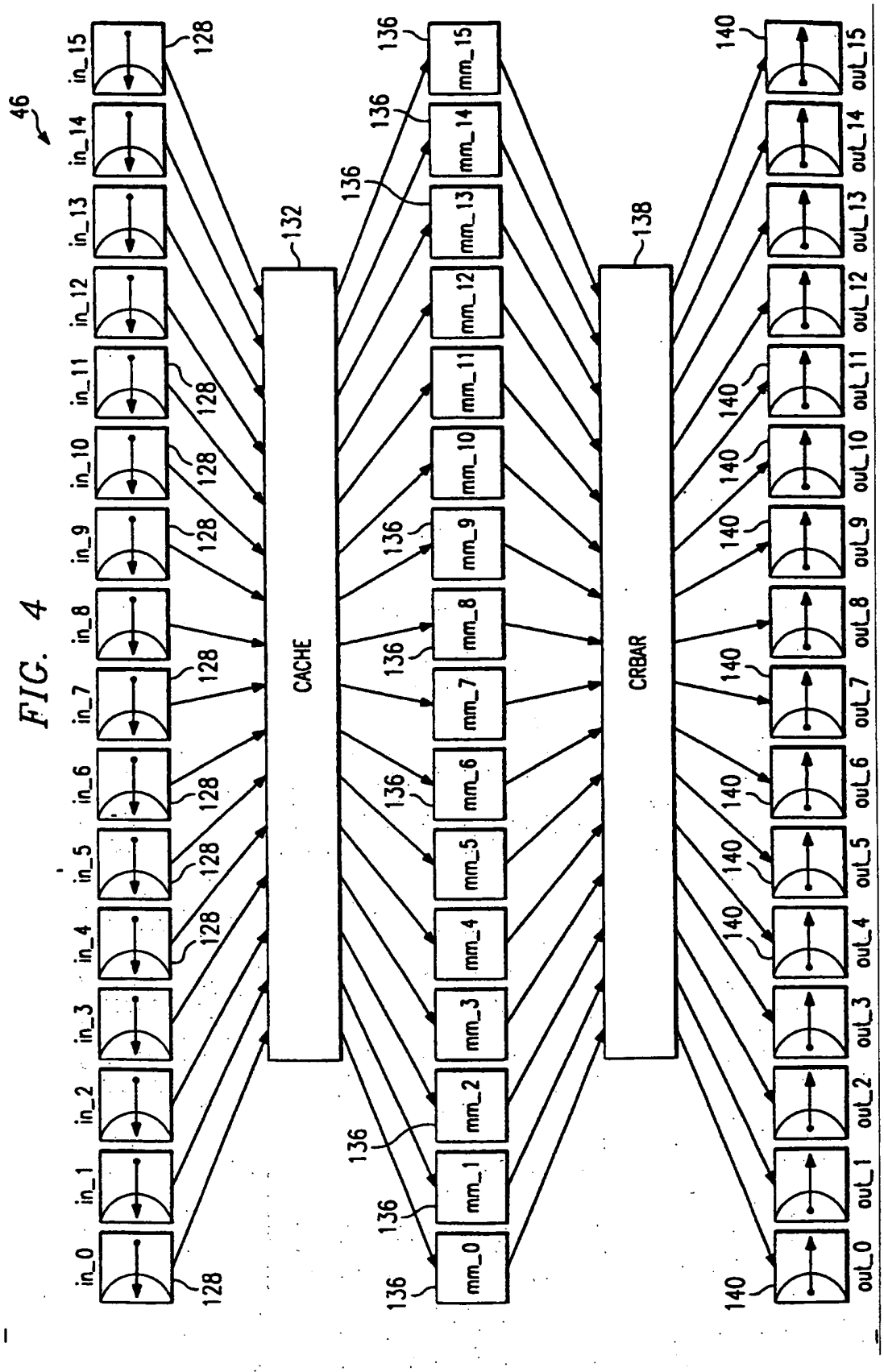
46
in_0
in_1
in_2
in_3
in_4
in_5
in_6
in_7
in_8
in_9
in_10
in_11
in_12
in_13
in_14
in_15
128
CACHE
132
mm_0
mm_1
mm_2
mm_3
mm_4
mm_5
mm_6
mm_7
mm_8
mm_9
mm_10
mm_11
mm_12
mm_13
mm_14
mm_15
136
CRBAR
138
out_0
out_1
out_2
out_3
out_4
out_5
out_6
out_7
out_8
out_9
out_10
out_11
out_12
out_13
out_14
out_15
140

FIG. 4

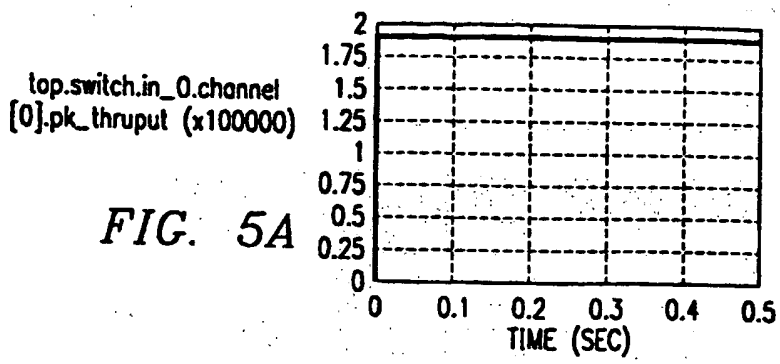

FIG. 5A

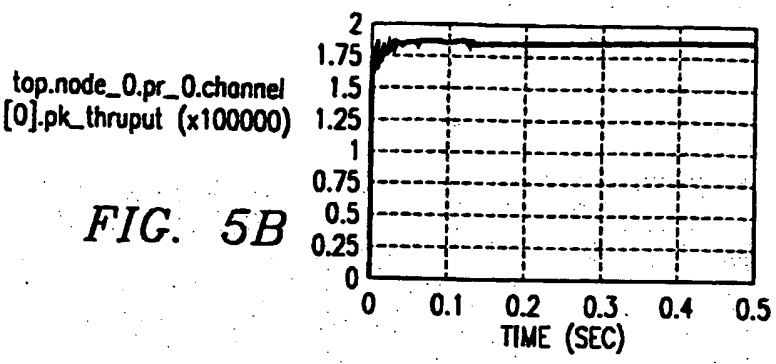

FIG. 5B

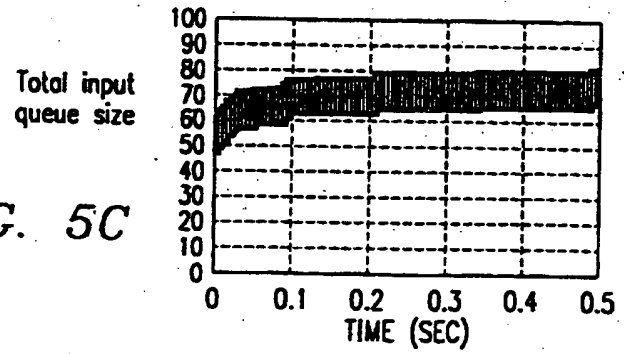

FIG. 5C

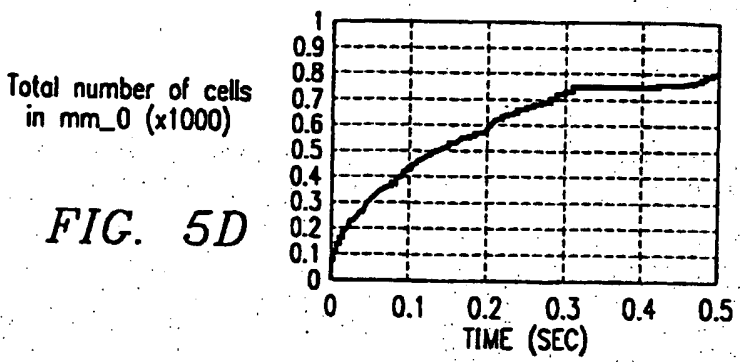

FIG. 5D

European Patent Office

EUROPEAN SEARCH REPORT

Application Number
EP 98 20 1768

DOCUMENTS CONSIDERED TO BE RELEVANT

Category	Citation of document with indication, where appropriate, of relevant passages	Relevant to claim	CLASSIFICATION OF THE APPLICATION (Int.Cl.6)
A	YAMANAKA H ET AL: "622 MB/S 8 X 8 SHARED MULTIBUFFER ATM SWITCH WITH HIERARCHICAL QUEUEING AND MULTICAST FUNCTIONS" PROCEEDINGS OF THE GLOBAL TELECOMMUNICATIONS CONFERENCE (GLOBECOM), HOUSTON, NOV. 29 - DEC. 2, 1993, vol. VOL. 3, no. -, 29 November 1993, pages 1488-1495, XP000436062 INSTITUTE OF ELECTRICAL AND ELECTRONICS ENGINEERS * section 2 *	1,2	H04L12/56 H04Q11/04
A	HIROMI NOTANI ET AL: "AN 8X8 ATM SWITCH LSI WITH SHARED MULTI-BUFFER ARCHITECTURE" PROCEEDINGS OF THE SYMPOSIUM ON VLSI CIRCUITS, SEATTLE, JUNE 4 - 6, 1992, no. -, 4 June 1992, pages 74-75, XP000342494 INSTITUTE OF ELECTRICAL AND ELECTRONICS ENGINEERS * section II *	1,2	TECHNICAL FIELDS SEARCHED (Int.Cl.6)
A	US 5 440 546 A (BIANCHINI JR RONALD P ET AL) 8 August 1995 * abstract; claim 1 *	1,2	H04L H04Q

The present search report has been drawn up for all claims

Place of search	Date of completion of the search	Examiner
THE HAGUE	7 September 1998	Lindner, A

CATEGORY OF CITED DOCUMENTS

X : particularly relevant if taken alone
Y : particularly relevant if combined with another document of the same category
A : technological background
O : non-written disclosure
P : intermediate document

T : theory or principle underlying the invention
E : earlier patent document, but published on, or after the filing date
D : document cited in the application
L : document cited for other reasons

& : member of the same patent family, corresponding document

EPO FORM 1503 03.82 (P04C01)